Practical statistical methods for call centres with a case study addressing urgent medical care delivery

David A. Wooff and S. Grace Stirling*

May 30th 2012

Abstract

Outside normal working hours, UK citizens requiring advice or treatment from a doctor must telephone a local out-of-hours call centre. Each call enters a queue and is subsequently assessed by a call-handler and classed as urgent or routine. Calls then enter another queue to await a response from a qualified doctor. A doctor telephones the patient to offer advice, or may arrange a home or centre visit.

From the patient's perspective, it is necessary to provide care in a timely fashion. From the perspective of Government funding of health care, resources are limited and certain lengths of queue are deemed acceptable. Typically, call centres must meet performance targets which relate principally to delays suffered by patients, as well as to quality of advice and treatment. As such, staff-loads must be carefully planned.

Planning needs accurate forecasts of incoming call volumes. These vary by hour, day, and season, and must account for calendar effects such as Christmas. We demonstrate the construction of models to predict interday and intraday call volumes. We show that simple models capture all the important features. We show how simulation models may then be used for resource allocation, uncertainty analysis, and staff scheduling. The data are details of call numbers and queue lengths from all parts of the patient-advice process for around five years, for a call centre based in Newcastle-upon-Tyne. There are around 350,000 complete cases in total. The methods are easily extended to other kinds of call centre. We describe the impact Swine flu had on call volumes in the summer of 2009, and our reactions to amend models in order to maintain forecast quality.

Keywords: Call-centre forecasting; prediction interval; daily arrival pattern; nonhomogenous Poisson process; patient queue; nurse scheduling.

1 Introduction

Our interest is in forecasting for call centres, and in particular out-of-hours call centres (OOHCC) which deal with patient requests for medical advice outside normal working hours. For such applications we need explicitly to take into account that demand varies strongly by day – because of a number of calendar effects, and by hour – because the call centre only operates at certain times and because demand varies with peaks typically during the early evening for weekdays and mornings at weekends. We refer to the forecasting of volume of calls expected each day as *interday* forecasting, and the pattern of calls arriving within the day as *intraday* forecasting. Other kinds of call centre may exhibit different interday and intraday arrival patterns; however the methodology we suggest should cope easily with other scenarios, without detailed modification.

The nature of medical emergency call-centre handling is that staffing appropriately is the key consideration. First, callers are patients in difficulty and there is an obligation to treat them as soon as possible, within reason. Typically there are specified guidelines on the timeliness of such treatment to which the call-centre needs to adhere. Secondly, patients are treated by qualified physicians. This is a very costly resource. The call-centre must therefore arrange staffing levels that are sufficient to meet demand subject to specified constraints, but must avoid prohibitively expensive idle capacity.

*Durham University, Department of Mathematical Sciences, Stockton Road, Durham DH1 3LE, UK. Email: d.a.wooff@durham.ac.uk

Interday and intraday forecasting can be correlated and as such, are generally modelled together Shen and Huang (2005, 2008); Weinberg et al. (2007). Avramidis et al. Avramidis et al. (2004) investigate the property that there is correlation between arrival counts during parts of the day, as well as dependency between arrival counts on successive days. For our purposes, interday and intraday forecasting need to be considered in some way separately. Interday forecasts were required to cover a volume of calls from midnight to midnight, as required by service commissioners to illustrate activity levels. Intraday forecasts run over staff shifts and are necessary to help with staffing levels at particular times.

Shen and Huang Shen and Huang (2005, 2008) apply an auto-regressive model incorporating an additional day-of-week effect to forecast interday rates. This lacks any seasonal or holiday component; for example, flu has a higher prevalence in winter than summer and demand will be higher then. Patient case-mix over holidays will also have an effect, for example patients seeking repeat prescriptions of routine medication may find their local clinic closed for a holiday period. We thus need a framework that will allow particular modelling of holidays and seasonal effects.

Weinberg et al. Weinberg et al. (2007) build a multiplicative Gaussian model for forecasting call centre arrival rates. They consider a nonhomogenous Poisson model modelling intraday and interday demand separately before combining them in a multiplicative model. The dataset used in the study was an eight-month period between March and October 2003, and again, seasonality and holiday effects were not included. The dataset in this case spanned over four years, and seasonality and holiday effects could be visualized quite easily, allowing a sensible model to be built to capture them.

Alternative approaches to modelling interday and intraday call centre demand are well documented in Shen and Haung's paper Shen and Huang (2008), which compares competing models by applying them to a case study. These include a linear model with day of week-and-time-of-day effects (referred to as the *historical average* approach), the Bayesian Gaussian model Weinberg et al. (2007), as well as multiplicative and additive models. Here they found their model performed well against the other methods, however it is still not adequate for application here without some seasonality and holiday components.

The U.S. Census Bureau's software package for seasonal adjustment, X-12-ARIMA, may also be considered, see for example Ladiray and Quenneville (2001). On testing this software, we found that we could not adapt it to forecast to the level that we wanted - daily, and with calls forecast for each 30-minute arrival period.

In studying and testing potential methods, we came to the conclusion that none serve very well. Some are statistically elegant, but are so narrowly focused to specific characteristics of an application scenario that it becomes difficult or impossible to adapt them to other scenarios. Our requirement is for accurate (because lives may be at stake) and adaptable (because calendar effects have strong implications for arrivals) methods. For our OOHCC scenario, we had the advantage of historical information for a small number of years. As such, fitting simple models to historical data is a natural starting point. As we shall see, one of the lessons we draw from this work is that simple models, thoughtfully applied, offer excellent performance. This is in line with Hand's Hand (2006) view that

> the improvements attributed to the more advanced and recent developments are small, and that aspects of real practical problems often render such small differences irrelevant, or even unreal, so that the gains reported on theoretical grounds, or on empirical comparisons from simulated or even real data sets, do not translate into real advantages in practice.

One might reasonably expect that forecasting technology is already routinely embedded within health services delivery. Alas, this is very far from the case. Few UK OOHCCs have credible statistically-based forecasting procedures. One reason for this may be a failure of academics to engage; another may be that existing methods overfit to particular details of a problem and so cannot easily be adapted. Academics tend to concentrate on publication, and on avant-garde methodology, rather than practical solutions.

In section 2 we describe the scenario for which we developed our methods. In section 3 we explain how we forecast call volumes arriving on specified days. In section 4 we explain how we forecast the pattern of arrival of calls during a specified day. In section 5 we discuss some features of operational use of the forecasts for planning and delivery of services.

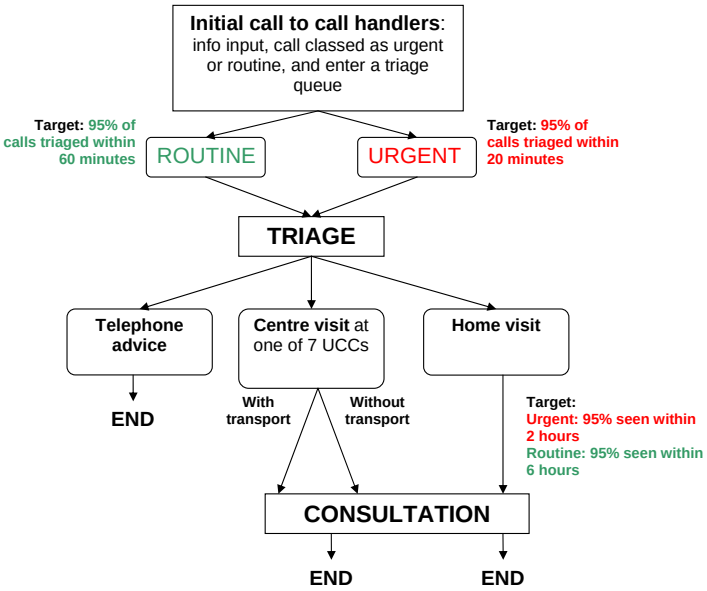

Figure 1: Flow chart of basic patient path through the OOH service

2 Out-of-hours medical call centre

In the UK, A General Practitioner (GP) OOHCC provides doctor cover between 18:30 and 08:00 on weekdays, and all day on weekends and bank (i.e. national) holidays. Companies providing this service are commissioned by local Primary Care Trusts. A patient rings the OOH number and reaches a non-clinical call-handler who records some details. Using specified urgency criteria, the patient's case is assessed at this stage as being *urgent* or *routine*. The call-handler will advise the patient that a GP will return their call, within a particular time frame. This process is termed *triage*. In most cases, one of three things will happen when the GP calls:

1. the patient's symptoms require advice over the telephone from a GP;
2. the GP decides that face-to-face consultation is necessary and makes an appointment for the patient at an Urgent Care Centre (UCC);
3. the GP arranges for a GP to visit the patient at their home.

This patient pathway is illustrated in Figure 1. There are key performance indicators (KPIs) for each stage in the pathway. We focus in this paper on the first stage, the triaging of calls, for which the targets are for Urgent (Routine) calls to be triaged within 20 (60) minutes. An OOHCC must score at least 95% in all KPIs to be classed as 'fully compliant' by their commissioners.

Nearly all cases are triaged by GPs based at the OOHCC. For centre visit consultations, a GP is based at an Urgent Care Centre with 10 minute appointment slots. When GPs have gaps in appointments, they can telephone in and help remotely with the triage queue, known as dipping into the queue. Therefore there is periodically extra resource available for triaging. For home visit consultations, a GP based at the OOHCC will be driven to the patient's home. These tend to be longer visits as they are generally more complicated cases. Afterwards, the GP will either travel back to the OOHCC, or to another home visit case. Whilst en route, GPs are encouraged to dip into the triage queue. Planning needs taking into account how many GPs will be away from telephone triaging and further logistics such as the number of drivers on shift.

For many call-centres, calls arrive directly as telephone calls. For our case study, as for some other medical emergency centres, the calls arriving are pre-screened by trained call handlers. There are clear

guidelines about which cases go through as genuine out-of-hours cases and whether they should be labelled as routine or urgent. Arrival rates are likely to be related to call handler behaviour; call handlers going on breaks at the same time, or working at different speeds. We could not observe or model this behaviour. Consequently, our focus is on pre-screened call arrival volumes.

2.1 Operational forecasting

Core to operational requirements of OOHCCs is appropriate staffing to balance patient care duties with achievement of KPI thresholds. Some of these questions are as follows. Which services are used more than others? How long do different services take? How many calls will arrive on any given day? When will calls arrive throughout a shift? How many GPs should be serving triage queues at any particular time? How many emergency vehicle drivers will be needed? Do all GPs exhibit the same service rates and level of quality? In the remainder of this paper, we consider the first, triage, stage of the patient path shown in Figure 1, and concentrate on two aspects: the volume of calls expected each day (Interday forecasting) and the pattern of calls arriving within the day (Intraday forecasting). Consideration of subsequent stages requires further statistical modelling, but employing arrival rates as the key ingredient.

As we noted earlier, we did not find an existing methodology to satisfy all our needs, in particular the need for good forecasts for highly situational time periods. As a first approach we thus decided to approach these two aspects separately. For the prediction of call volumes per day we used linear regression. This worked well as it incorporates dummy variables easily, making handling different holiday effects unproblematic. For the pattern of call arrivals through the day we explored a number of options before settling on a loess approximation to an assumed Poisson arrival rate with discrete portions of the day, as a feasible means of getting at the underlying nonhomogenous Poisson process.

3 Interday forecasting

3.1 Data collection

Out-of-hours cases were collected from January 1^{st} 2005, and these were used to build a predictive model to forecast calls per day. At the start of our work we had available over 200,000 telephone calls per day (CPD) for around 1,400 consecutive days. Figure 2 shows the dataset for CPD between January 1^{st} 2005 and June 30^{th} 2009. There is a day pattern, with low volumes for weekdays and higher volumes for weekends and bank holidays. There are high-variation peaks at particular times during the year. There are some cyclical and seasonal patterns.

3.2 Linear regression

We use the standard multiple linear regression setup Draper and Smith (1998). Thus, suppose that the response variable y represents the number of calls per day, with observed value y_i on day i, $i = 1, \ldots, n$ collected into the $n \times 1$ vector $\mathbf{Y}$. We fit a model with an intercept and $p-1$ covariates to be selected from those available. The design matrix is $\mathbf{X}$, an $n \times p$ matrix whose first column is a vector of ones and whose remaining columns are the vectors $\mathbf{x_1}, \ldots, \mathbf{x_{p-1}}$, where $\mathbf{x_j}$ is the $n \times 1$ vector of values for covariate j. The coefficient vector is the $p \times 1$ vector β, with β_1 giving the model intercept. The linear model is $\mathbf{Y} = \mathbf{X}\beta + \epsilon$, with ϵ an $n \times 1$ vector representing independent homoscedastic $N(0, \sigma^2)$ errors. The coefficients are estimated via least squares giving $\widehat{\beta}$, and these then generate fits as $\widehat{\mathbf{Y}} = \mathbf{X}\widehat{\beta}$, and residuals as $\widehat{\epsilon}_i = y_i - \widehat{y}_i$. Model adequacy is assessed *inter alia* through Wald-type tests on individual coefficients, F tests on the overall fit, and residual and influence diagrams.

3.3 Stepwise regression

The response variable, y, may depend on a number of covariates and their possible interactions. In order to explore suitable models, we employed stepwise regression Draper and Smith (1998), using a combination of forward selection and backward elimination and bearing in mind the plausibility of the resultant models. The covariates we considered are shown in Table 1. In the UK, national holidays are known as bank

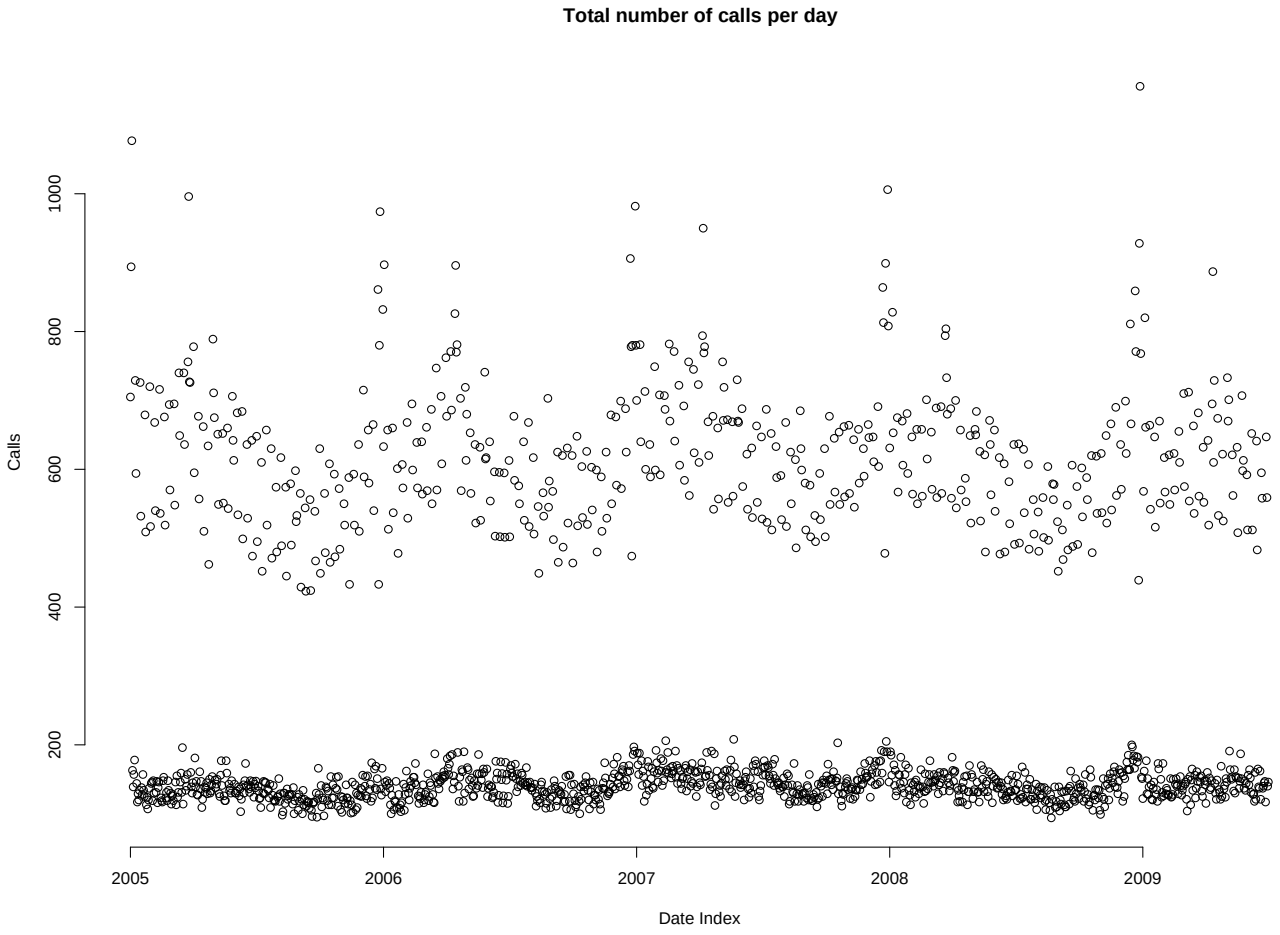

Figure 2: Calls per day between 01/01/2005 and 30/06/2009

Table 1: List of candidate factors thought to be useful in predicting daily call volumes, CPD. Terms shown in a bold font term were included in the final model.

Factor	Rationale
dow	Day of week: OOH service runs for 14.5 hours on weekdays and 24 hours on weekends
month	Month: Suspect summer months are quieter than winter months
year	Year: Are there differences in demand between years?
time	Chronological date: Are there changes in demand over time?
easterbh	Easter weekend: Busier than normal Fridays-Mondays in March and April.
easterfri	Good Fridays: Busier than normal Fridays in March and April.
eastersat	Easter Saturday: Busier than normal Saturdays in March and April.
eastersun	Easter Sunday: Busier than normal Sundays in March and April.
eastermon	Easter Monday: Busier than normal Mondays in March and April.
eastmarch	Easter holidays in March: Expect Easters in March to be busier than April as the weather tends to be colder.
otherbh	Bank Holiday weekends in May and August: We expect these days to be busier than normal Saturdays-Mondays in May and August.
bhsat	Bank Holiday Saturdays: Busier than normal Saturdays in May and August.
bhsun	Bank Holiday Sundays: Busier than normal Sundays in May and August.
bhmon	Bank Holiday Mondays: Busier than normal Mondays in May and August.
maybhsat	Bank Holiday Saturdays in May: Busier than normal Saturdays in May.
maybhsun	Bank Holiday Sundays in May: Busier than normal Sundays in May.
maybhmon	Bank Holiday Mondays in May: Busier than normal Mondays in May.
augbhsat	Bank Holiday Saturdays in August: Busier than normal Saturdays in August.
augbhsun	Bank Holiday Sundays in August: Busier than normal Sundays in August.
augbhmon	Bank Holiday Mondays in August: Busier than normal Mondays in August.
maybh1	First May Bank Holiday weekend: Busier than other Bank Holiday weekends.
xmaswd	Christmas Day falls on a weekday: Busier than a normal weekday in December.
xmaswe	Christmas Day falls on a weekend: Different to a normal weekend in December.
xmasbhwd	Christmas holidays on weekdays: Public holidays on weekdays will be busier than normal weekdays in December and January.
xmasbhwe	Christmas holidays on weekends: Busier than normal weekends in December and January
nydwd	New Years Day falls on a weekday: Busier than a normal weekday in January.
nydwe	New Years Day falls on a weekend: Different to a normal weekend in January.
xmas4day1	Days following Christmas when it's a four-day holiday cycle: Unknown influence of the difference between a four-day Christmas holiday versus a two-day Christmas holiday.

holidays. Although the full data set is very large, we need to take into account that the number of possible interactions is also very large and we quickly exhaust degrees of freedom – and computational power – in fitting full models. We need also to take into account that very large sample sizes lead inevitably to tiny, and often misleadingly small, p-values. Therefore our stepwise modelling will depend partly on checking for statistical significance, partly on plausibility, and partly on checking residuals to see whether enhancing a model has practical benefit. In arriving at such models we are guided by the loss or gain in variance explained using partial F-tests whilst paying attention to model parsimony using criteria such as Akaike's information criterion and Mallows' C_p. Box-Wetz methods Draper and Smith (1998) may be used to check on whether a model is worthwhile. We bear in mind that stepwise regression modelling is ad hoc, and knowing where to stop is often a subjective judgement requiring statistical experience.

3.4 Predicting daily call arrival volumes

Table 1 shows the covariates we took into account, with some constructed from others. We explain in the table our reasons for considering them. After applying stepwise regression, we settled on a model which included the terms marked in bold font in Table 1, plus interactions between *month* and *we*, *month* and *time*, and *easterbh* and *time*. This model has a large number of parameters, for example there are six coefficients for day of week, *dow*, with Mondays as baseline; and 11 coefficients for month, *month*, with January as baseline. There is some confounding to beware. For example we must be careful in interpreting *time*, a linear effect intended to capture a steady increase or decrease in demand, with *year*, which allows essentially a different baseline for each calendar year.

We omit from this account details of the very many models explored, except to illustrate some of the considerations. For example, we found that we needed to include explicit terms for each day of the Easter holiday. On the other hand, May and August bank holiday weekends may be handled through separate dummy variables for Saturday, Sunday, and Monday effects plus a dummy variable to explain that May bank holiday call volumes tend to be higher than those for August. Generally, call volumes in the summer tend to be lower than in the winter, possibly because the North-East regional population drops as people are away on holiday and possibly because of generally lower prevalence of illness. Public holidays tend to be significantly busier; Christmas being the busiest, followed by Easter, and May and August Bank Holidays respectively. Details of models tried and other aspects of the fitting processes may be found in Stirling (2011), which also contains coefficient values, results of tests, and so forth.

The standardized residual plot shown in Figure 3, is mostly highly satisfactory. There are some high residuals, but almost all relate to the Christmas and New Year periods. One reason for this is that we have very limited data available for these holiday periods, once it is recognized that the particular location of Christmas day within the week has a strong effect on call patterns. For example, it is thought that call volumes can be exceptionally high on the Saturday immediately following Christmas day whenever this occurs. Partly because of lack of data, our model cannot yet capture well all possible Christmas and New Year configurations. As such, the resulting predictions for such periods have larger variability. These will be better resolved as fresh data arrives. Historically, the partner OOHCC has very high uncertainty concerning call volume patterns over the Christmas period, with the result that shift planning frequently is inappropriate and KPIs are missed by very wide margins.

Such linear models routinely provide confidence intervals on the mean response and prediction intervals for particular forecasts. Thus, one output to call-centre managers is a forecast of call volume for any day in the year, together with a prediction interval. As an example, Table 2 shows the call volume forecast and prediction interval for the days of the second UK May Bank Holiday in 2009. Generally, the forecasts using the model have been highly trustworthy but with occasional poor forecasts occurring particularly at Christmas. Forecasting over the Christmas 2008 period showed a large underfit, and there was overfitting for Christmas 2009. This can partly be explained by operational reasons such as telecommunications failure and unexpected staff absence.

3.5 Model updating and validation

In operational terms, we began applying these ideas with around 2-3 years data, with initial results available from linear modelling for the Christmas 2008 period, as this is the time of year for which forecasts are most

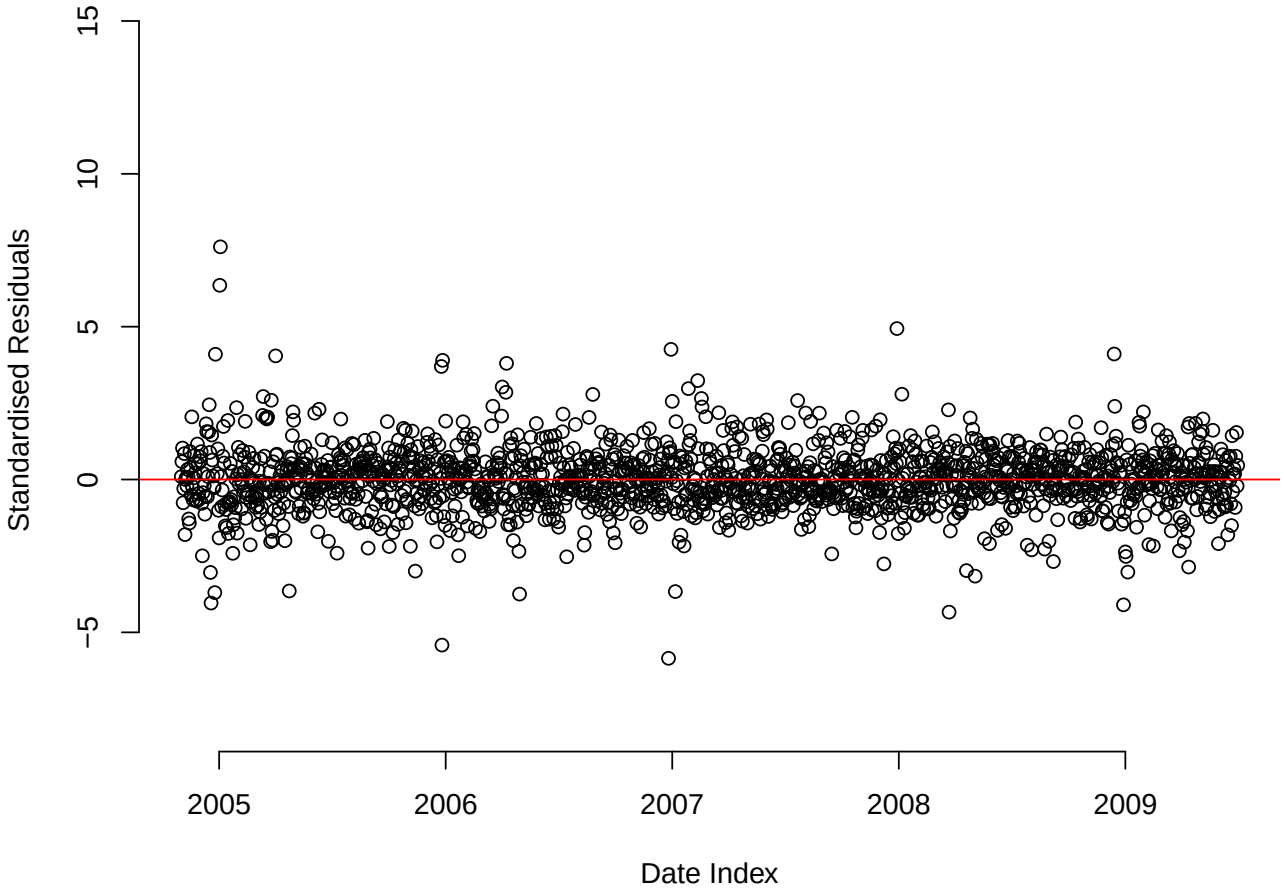

Figure 3: Standardized residual plot for the model for the period 01/11/2004 to 30/06/2009

Table 2: Forecasts, prediction intervals and observations for May Bank Holiday 2009

Date	Forecast	95% Prediction Interval	Observed
23/05/2009 (Sat)	726	(649,802)	713
24/05/2009 (Sun)	652	(575,729)	606
25/05/2009 (Mon)	649	(571,721)	613

keenly sought. As time progressed, we periodically updated our models. Typically this means retaining existing model terms, with slighter sharper estimates reflecting more data and higher certainty. As much more data arrives it is possible to add new terms to the model, with the interpretation that more data allows us to pick up previously undetectable influences. Each forecast is compared to the corresponding observation as it arrives. This allows us to diagnose potential model inadequacy, but usually investigation reveals an operational cause for the discrepancy.

We were fortunate to obtain access to an earlier data set covering two months from 2004, partial in the sense of the full patient pathway shown in Figure 1, but complete in terms of call arrivals. Therefore we were able to validate our model based on post-2005 data to predict 2004 outcomes. We found no obvious residual or temporal discrepancies, and comparison of residuals from the model data set and the validation data set revealed no significant differences.

3.6 Adapting the model to other scenarios

Linear modelling employing dummy variables for calendar effects works very well as a pragmatic solution and is easy to implement. As a second example, while we were working on OOHCC forecasting, we were approached by Sustrans, a UK charity concerned with encouraging environmentally friendly travel. The data consisted of counts of cycles in 6 UK cities for 2005-2009. For each city, there are several cycle counters which automatically record cycle traffic. In total there were 111 locations with counters. Interest is in the seasonal pattern of cycle use, together with the implications of covariates such as route type, amount of street lighting, proximity of school, and so forth. The number of observations is approximately 125,000. To analyze the data, we employed essentially the same methodology to assess seasonal calendar effects for each counter separately.

The standard linear multiple regression model was tried, both on the raw counts and taking square roots of counts as theory suggests these have a more Gaussian distribution. These gave reasonable fits, with typically around 85% variation explained, depending on location. Unlike the call-centre counts, these cycle counts can be much smaller and so it is more satisfactory to model the counts directly as Poisson. As such, a generalized linear model was also fitted using Poisson regression (log link). These also gave good fits but tended to suggest overdispersion. Consequently, negative binomial regression models were also tried, and this was the final model accepted. In checking assumptions and model fit, there were no worrying discrepancies. There were no obvious missing seasonal or other terms, and relatively few very large standardized residuals. Results from the analysis were ultimately used by the UK Department of Transport to address policy.

3.7 Unexpected shocks to the system

Managers may need to adapt models when there are shocks to the system, typically a new and unexpected external factor which must be taken into account. An example is provided by the Swine flu panic of 2009, which had a major impact on call volumes coming into the OOHCC. Headlines concerning Swine flu began to be reported towards the end of April 2009. By July 2009 there was a massive surge in activity levels for healthcare providers across the world; completely uncharacteristic and unprecedented. Patients with even quite mild flu symptoms began to call the OOHCC for advice where previously they would have managed the symptoms themselves. Part of our task with our parter institution, was urgently to model the implications of the epidemic on call volumes. No methodology can handle such shocks straightforwardly. However, one advantage of the regression approach is that extra model terms can be added and explored quite easily. The basic model already provided forecasts of call volumes in the absence of Swine flu; therefore the excess between observations and daily forecasts can be assumed to be due to extra calls generated by the panic, plus random variation. This provides a sequence of pseudo-observations which may be modelled explicitly to capture the shape of the epidemic process – we chose a simple moving average. Once a reasonable model for the epidemic has been found, it can be added to the basic model for the duration of the epidemic. This strategy worked very well in practice, but requires the assumption that the epidemic is not confounded with parallel changes in underlying healthcare needs.

Table 3: Dates in August used for the production of forecasts for 'Saturdays in August'

Year	Dates	Year	Dates
2005	6^{th}, 13^{th}, 20^{th}	2008	2^{nd}, 9^{th}, 16^{th}, 30^{th}
2006	5^{th}, 12^{th}, 19^{th}	2009	1^{st}, 8^{th}, 15^{th}, 22^{nd}
2007	4^{th}, 11^{th}, 18^{th}	2010	7^{th}, 14^{th}, 21^{st}

4 Intraday forecasting

Intraday forecasting is necessary to plan staffing to cope with an expected pattern of call arrivals during the day. Call arrival rates are not constant, but exhibit peaks and troughs in demand, depending on the time of day. Our aim is to produce a forecast of the arrival rate for every 30 minute period of every day in the year, taking into account all calendar effects. We choose to simplify by discretizing to 30-minute periods rather than modelling the corresponding continuous time problem, at least partly because this suits real-world practicalities of arranging work periods.

To establish forecasts, we employ a version of the dataset discussed earlier, but extended to take into account fresh data arriving up to September 2010. Between 01/01/2005 and 01/09/2010 we have a total of 594,980 individual out-of-hours observations (individual phone calls) over the various days, times and seasons. For operational reasons, the forecasting periods are defined by shifts, for example Monday 18:00:00 to Tuesday 07:59:59 and Saturday 08:00:00 to Sunday 07:59:59. Weekdays have 28 non-overlapping 30-minute time periods. Weekend days and bank holidays have 48 time periods. As with the calls per day model, we assess bank holidays explicitly. We take into account that days adjoining bank holidays, such as the Saturday preceding August bank holiday, may need special treatment. The simplest way of proceeding is to assume that a reasonable forecast for arrival patterns within a given day can be generated by averaging in some way the patterns of arrivals from all the other days which we deem to be sufficiently similar.

There are three key challenges. The first is in deciding which days are similar. The second is to find ways to average the patterns that we can see, and the third is to provide some kind of uncertainty envelope for them. Guided by modelling of interday volumes, we classify days by weekday, month, and special calendar effects such as bank holiday. This allows us to capture variation in pattern of arrival due to day of week, which we know to be substantial. By specifying month, we are allowing for some general seasonal differences. For example, we might expect in summer some calls in early evening due to sporting injuries, an activity probably reduced in winter. Such a classification is guided by discussion with representatives from the organization.

Figure 4 shows observations for Saturdays in August, with the dates of these observations detailed in Table 3. We exclude Saturdays during a bank holiday period, as we consider these separately. We use only these observations to make forecasts for Saturdays in August 2011. There are a total of 960 observations; 20 observations for each of the 48 time points. There is a clear pattern with a sharp increase between 08:00 and 09:29, a steady decrease between 09:30 and 12:59, a slower decrease between 13:00 and 23:59, and a quieter period overnight between 00:00 and 06:59. This corresponds closely with what call centre managers know intuitively. Note that there is quite significant variation in the number of calls arriving into each time slot, and this variation does not seem to be constant.

We may fit the arrival patterns using a number of different methods. One would be to employ classical time series. Another would be to use functional data analytic methods Ramsey and Silverman (2005), which might reveal interesting features through the functional principal components. Instead, we begin with simple possibilities.

4.1 Locally Weighted Regression

Locally weighted regression is a multivariate smoothing procedure which fits a regression surface to a set of data points Cleveland (1979) Cleveland and Devlin (1988). Let y_i, $i = 1, ..., n$ be measurements of a response variable and let $x_i = (x_{i1}, ..., x_{ip})$, $i = 1, ..., n$ be the corresponding vector of measurements of p

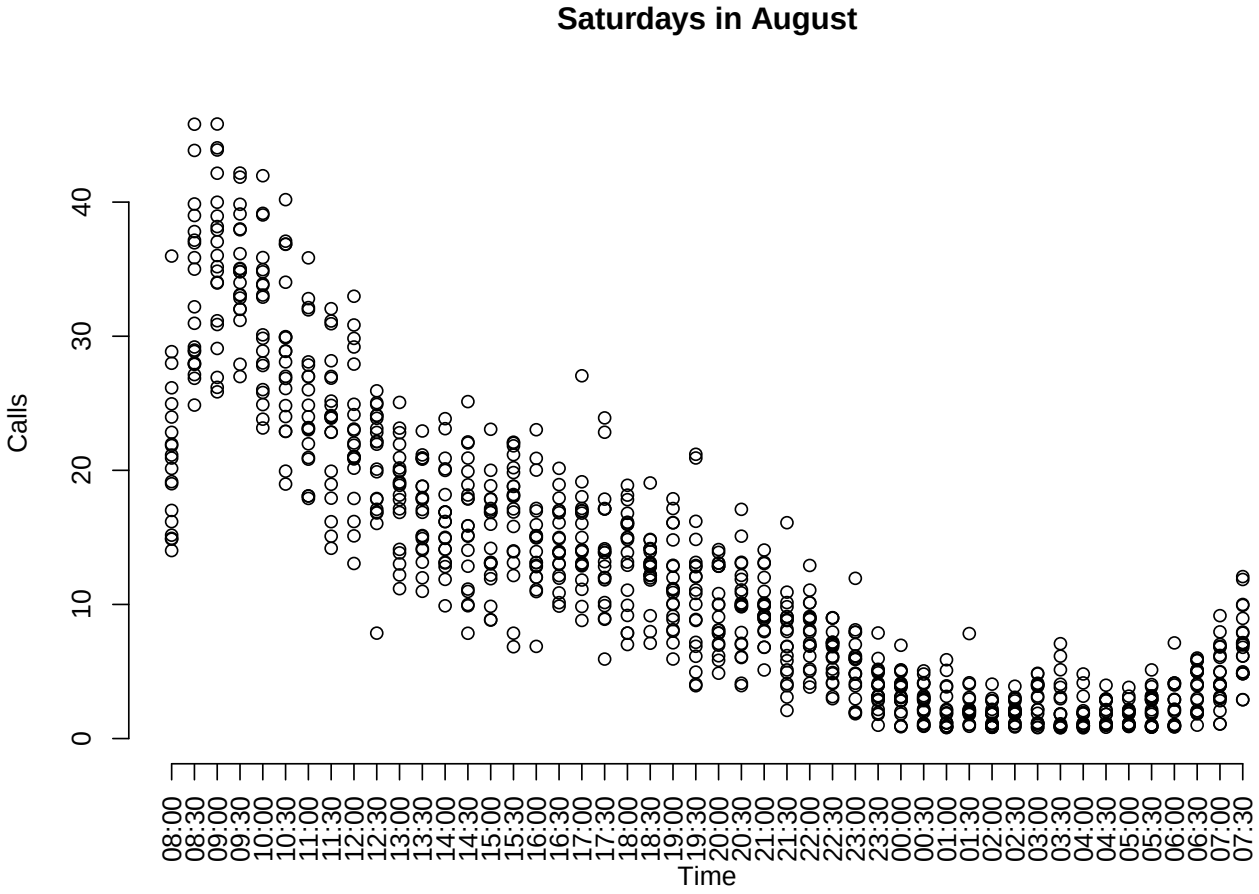

Figure 4: Previous half-hourly observations with random variation

predictors. The relationship between the response and predictors is

$$y_i = g(x_i) + \epsilon_i,$$

where we assume $\epsilon_i \sim N(0, \sigma^2)$ are random errors and g is a smooth function of the predictors. Locally weighted regression provides an estimate $\hat{g}(x)$ at each value x_j, $(j = 1, ..., p)$. The estimate of g at x_j is obtained from a neighbourhood of points, weighted according to their distance from x_j. Points close to x_j have a large weight, and points far away from x_j have a small weight. To carry out locally weighted regression we need a distance function, ρ, a weight function, W, and the specification of a neighbourhood size. For further details, see for example Chambers and Hastie (1993). To make the fits we employ the `loess` function provided in the **R** package R Development Core Team (2009). Fitting is performed locally. Because of the curvature evident in Figure 4, we use locally quadratic fitting by least squares.

The variation of the overall pattern shows some, but not much, unusual behaviour. The observations at each time period seem roughly evenly spread with possible exceptions early in the morning, around 12:30, and 17:00. These correspond to opening times and lunchtime and end-of-work times. We thus begin with the assumption that errors are Gaussian. In what follows, the span is always less than 1, so the neighbourhood includes a proportion of the points, and these have tricubic weighting.

For a fixed day in the year, the response variable is the volume of calls per time period, which we will denote by C (arranged in a $p \times n$ matrix), and the predictor is the half-hourly time period T (arranged in a $p \times 1$ matrix), which we may arrange as the $p \times (n + 1)$ matrix:

$$A_{p,n+1} = \begin{pmatrix} t_1 & c_{1,1} & \cdots & c_{1,n} \\ t_2 & c_{2,1} & \cdots & c_{2,n} \\ \vdots & \vdots & \ddots & \vdots \\ t_p & c_{p,1} & \cdots & c_{p,n} \end{pmatrix}$$

where $p = 48$, $n = 20$, $c_i = (c_{1,i}, ..., c_{48,i})$, $i = 1, ..., 20$, are the previous observations for day i, (the response), and $t = (t_1, .., t_{48})$ are the 30 minute time periods running from 08:00 on the Saturday until 07:59 the next Sunday. This matrix is principal input to the loess procedure, along with a span choice. A span of 0.15 was chosen because it worked well in practice, as shown in Figure 5. A higher span choice tended to underestimate the 08:00 peak. A lower choice results in the fit becoming too spiked as in Figure 6. Choosing a span of 0.15 for weekdays (14 hour periods) and weekends (24 hour periods) means the smoothing will be slightly different for weekdays and weekends. However, in practice this choice worked well to capture arrival rates, and was adopted in the partner organization.

4.2 Establishing a Prediction Interval

After establishing a reasonable method for predicting future patterns of call arrivals, we now need to place confidence bounds on these predictions. However, with loess as the fitting procedure there is no agreed way to do this. We may obtain a standard error for the fit, s_F, but there is no accepted standard error for a prediction, s_P. We now explore two possibilities for constructing prediction intervals.

4.2.1 Approximate t-based errors

First we adapt an idea from ordinary least squares (OLS) regression. Recall the relationship between the response and predictor; $y_i = g(x_i) + \epsilon_i$, where $\epsilon_i \sim N(0, \sigma^2)$, σ is unknown, and g is a smooth function. In OLS, when σ is unknown, we generate standard errors of the fitted values Draper and Smith (1998) using

$$s_F = \hat{\sigma}\sqrt{\frac{1}{n} + \frac{(x - \bar{x})^2}{s_{xx}}}, \tag{1}$$

where $\hat{\sigma}$ is the usual estimate of σ and $s_{xx} = \sum_i^n (x_i - \bar{x})^2$. These standard errors are used when generating approximate confidence intervals on the mean response, $\hat{y} \pm t_{1-\alpha/2} s_F$, assuming Normality of errors, where t is the t distribution with $n - p$ parameters, and α is the level of significance. When we are trying to

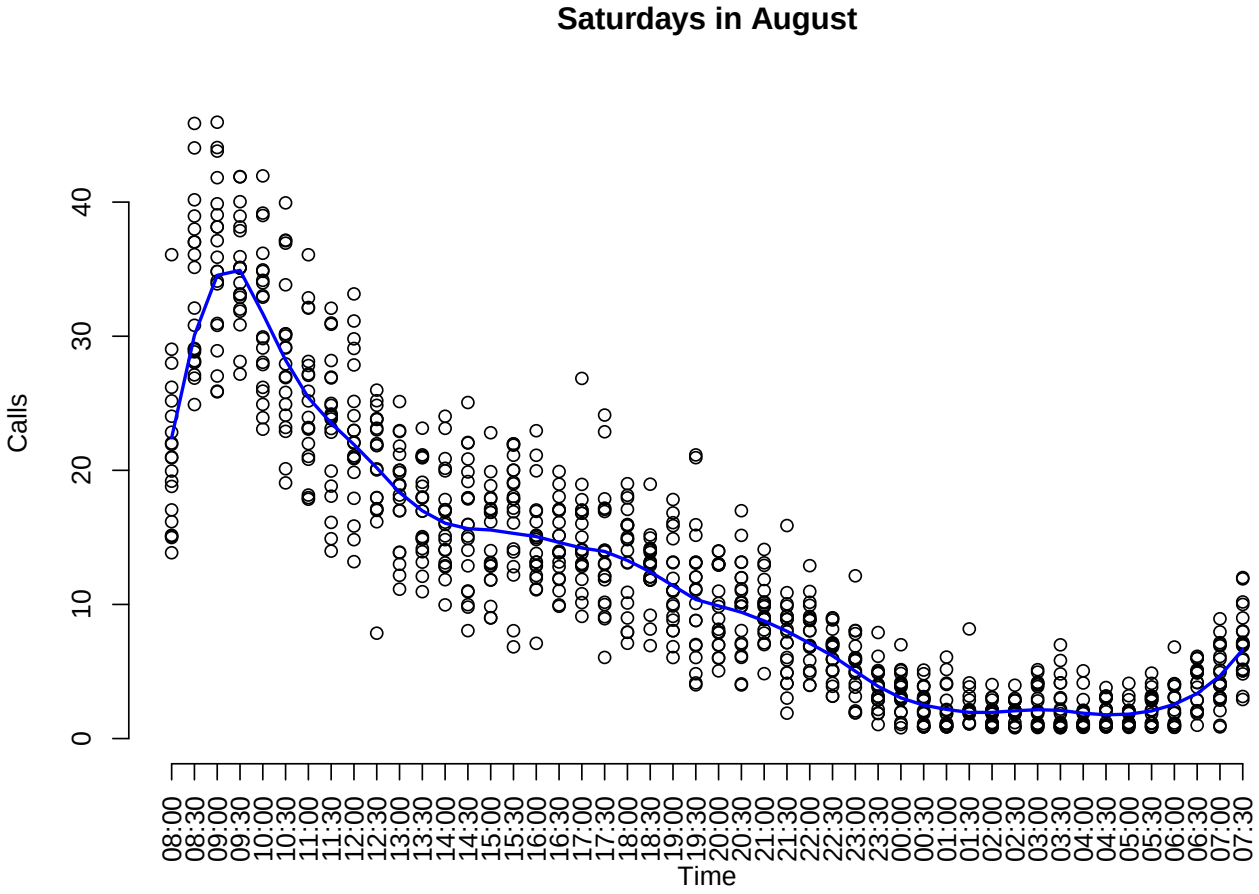

Figure 5: Previous half-hourly observations with random variation with loess fit using span=0.15

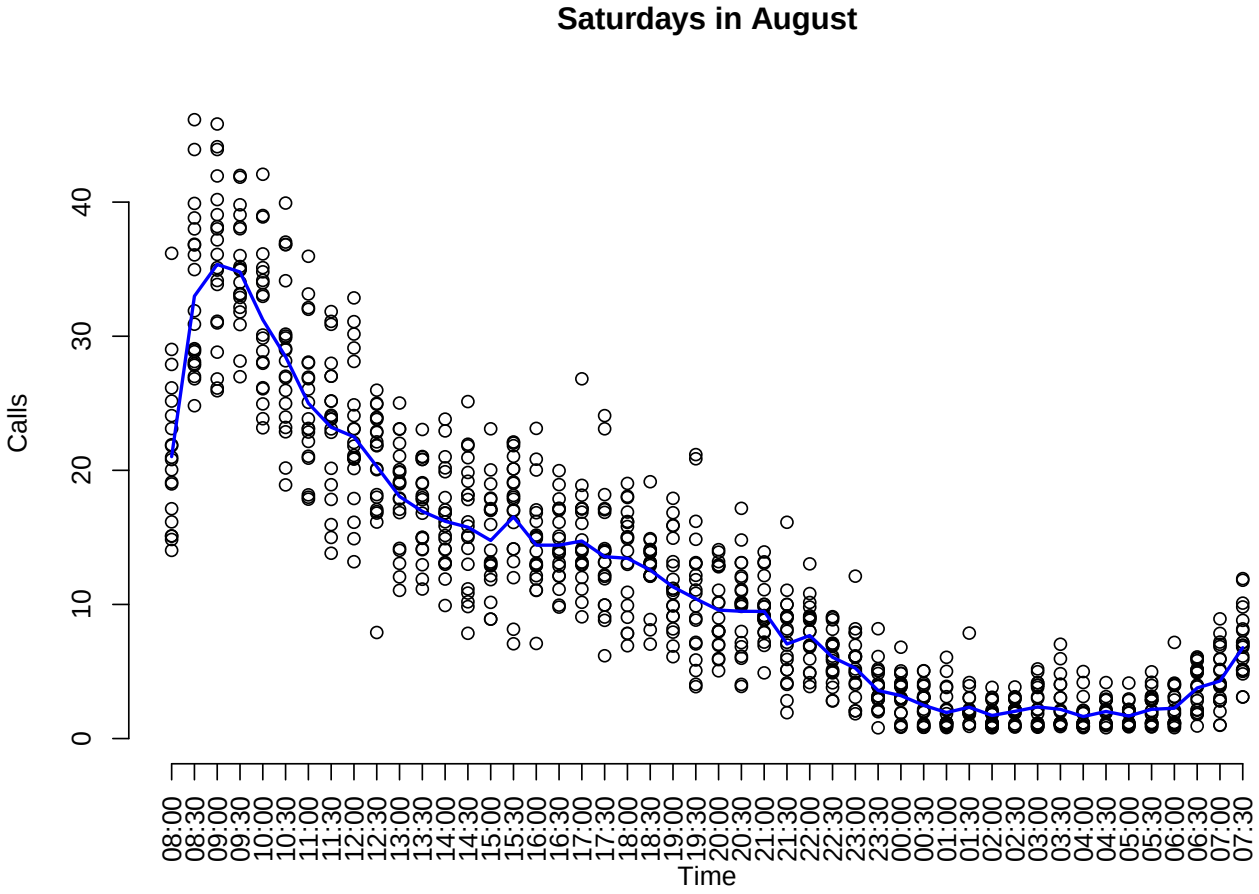

Figure 6: Previous half-hourly observations with random variation with loess fit using span=0.05

predict a new observation $y_0 = g(x_0) + \epsilon_0$, at a proposed value x_0, the corresponding prediction interval is based on the standard error for the prediction:

$$s_P = \hat{\sigma}\sqrt{1 + \frac{1}{n} + \frac{(x_0 - \bar{x})^2}{s_{xx}}}, \tag{2}$$

with prediction intervals generated by $\hat{y}_0 \pm t_{1-\alpha/2} s_P$. The fits are the same whether generating confidence or prediction intervals, but the prediction interval is wider than the confidence interval to account for uncertainty in the error at x_0 as well as uncertainty in the mean response. The standard errors are related as $s_P^2 = \hat{\sigma}^2 + s_F^2$. This suggests that if we have reasonable estimates from the loess fits for σ^2 and s_F, we can generate an approximate prediction error s_P. To do so, and assuming we wish to base intervals on the t_{n-p} distribution, we need to have some idea of the equivalent number of parameters p implicit within a loess fit. A by-product of loess fitting is l_p, representing the equivalent number of parameters used in the loess fit, and which depends on the span chosen for the fit Chambers and Hastie (1993). As such we base our intervals on the t_{n-l_p} distribution, taking $l_p = max(2, l_p)$.

In practice this method produced intervals which worked well over busy parts of the shift, but where the upper limit tended to be too generous for the overnight part of the shift. This is illustrated in Figure 7. This shows the previous observations for Saturdays in August, with the central line illustrating the loess fit, and the upper and lower lines giving the prediction interval. Also shown is a Normal quantile plot of the residuals (left) and a scatter plot of the standardized residuals (right). There is curvature in the quantile plot, suggesting that Normality of errors is an unsafe assumption. The standardized residuals appear not randomly scattered, suggesting model deficiency.

4.2.2 Approximate Poisson-based errors

Usual models for call arrivals start with a Poisson assumption for arrival rates and thus an exponential assumption for inter-arrival times. Indeed, a more sophisticated approach would be to take the arrival times as resulting from a continuous time nonhomogenous Poisson process, with intensities dependent on calendar covariates and time of day. This has been the approach taken elsewhere, for example Weinberg et al. (2007). However, we may make weaker assumptions based simply on the idea that numbers of arrivals within a given 30-minute time period are Poisson with rate λ_t, independently of non-overlapping intervals. Goodness-of-fit tests using counts of arrivals from several separate time periods for several days showed that this assumption is reasonably safe, though not perfect. We thus proceed under this assumption.

For our first method, we noted some uneven variation. When a variable is Poisson distributed, its square root is approximately normally distributed with expected value of about $\sqrt{\lambda}$ and with variance of about 1/4. This is an example of a variance-stabilizing transformation. This suggests an alternative strategy.

We begin by transforming to approximate Normality by taking $\tilde{c}_{tj} = \sqrt{c_{tj}}$, $j = 1, \ldots, n$, for each time period. This also helps to stabilize the variance. We then apply loess to these transformed values. This gives a predicted value for the mean response within that interval, which we take to be an estimate of $\sqrt{\lambda_t}$, where λ_t is the arrival rate for time interval t. The use of loess fitting assures that the arrival rates are reasonably smoothly related from one interval to the next. For each time interval we may now generate a forecast and a prediction interval for the number of calls arriving in that interval based upon a Poisson(λ_t) distribution. This provides a piecewise prediction interval for arrival rates for a specific (day*month) or calendar effect combination. The predicted mean arrival rates are then normalized so that the daily forecast total for intraday matches the prediction arising from the interday modelling. In practice, these give very good results with no substantial anomalies.

This is illustrated in Figure 8. The method has produced a plausible fit and a prediction interval which captures the heteroscedasticity far better. The fit to the observations appears superior. The Normal quantile plot shows no deficiencies, and the standardized residual plot has improved from Figure 7. There does appear to be pattern in the residuals on the overnight part of this shift, between 22:30 - 07:59. The pattern occurs because there is very little forecast activity and the residuals are due to discrepancies between small forecast non-integers and small integer observations.

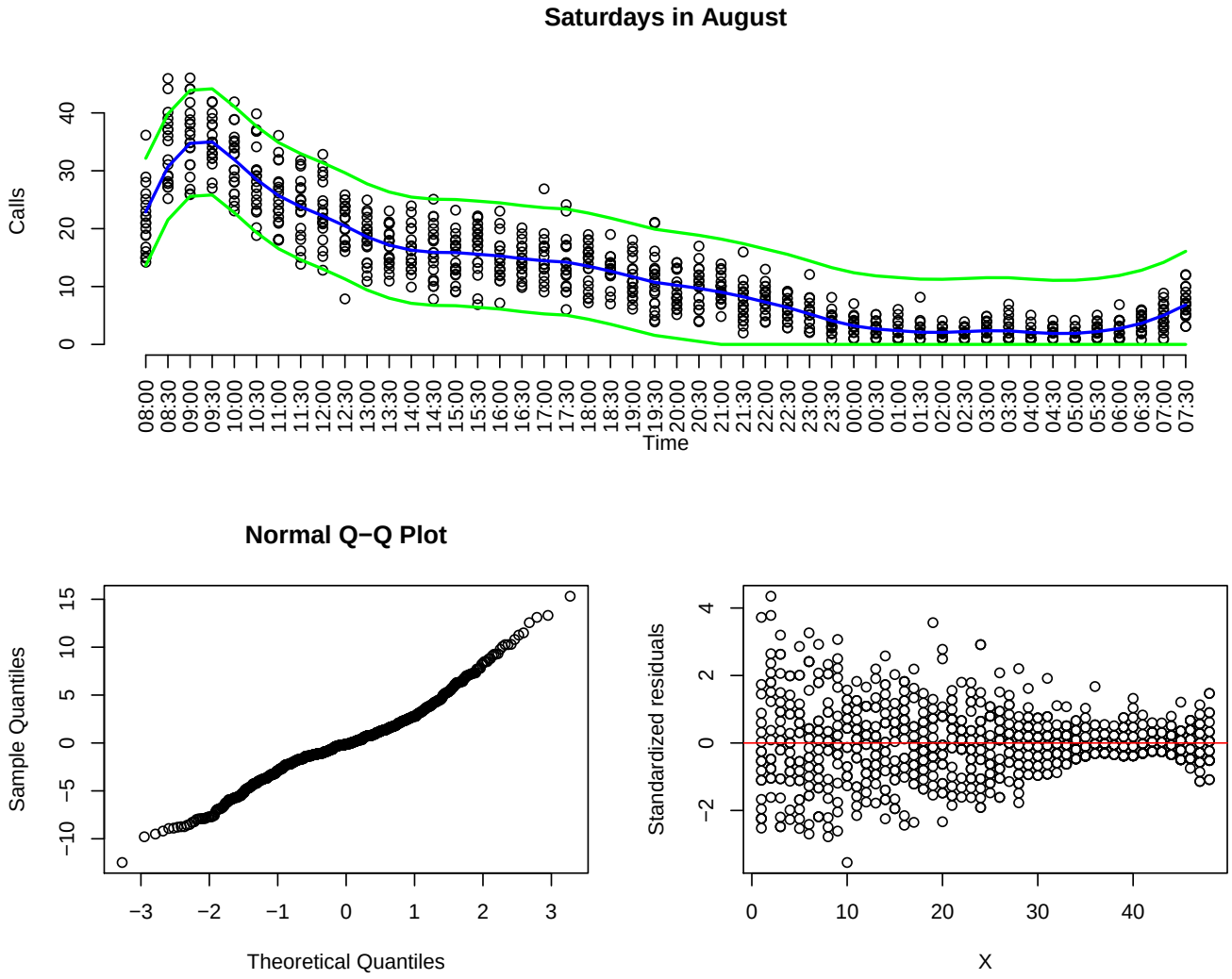

Figure 7: Observations with loess fit and t-based errors with 95% prediction interval.

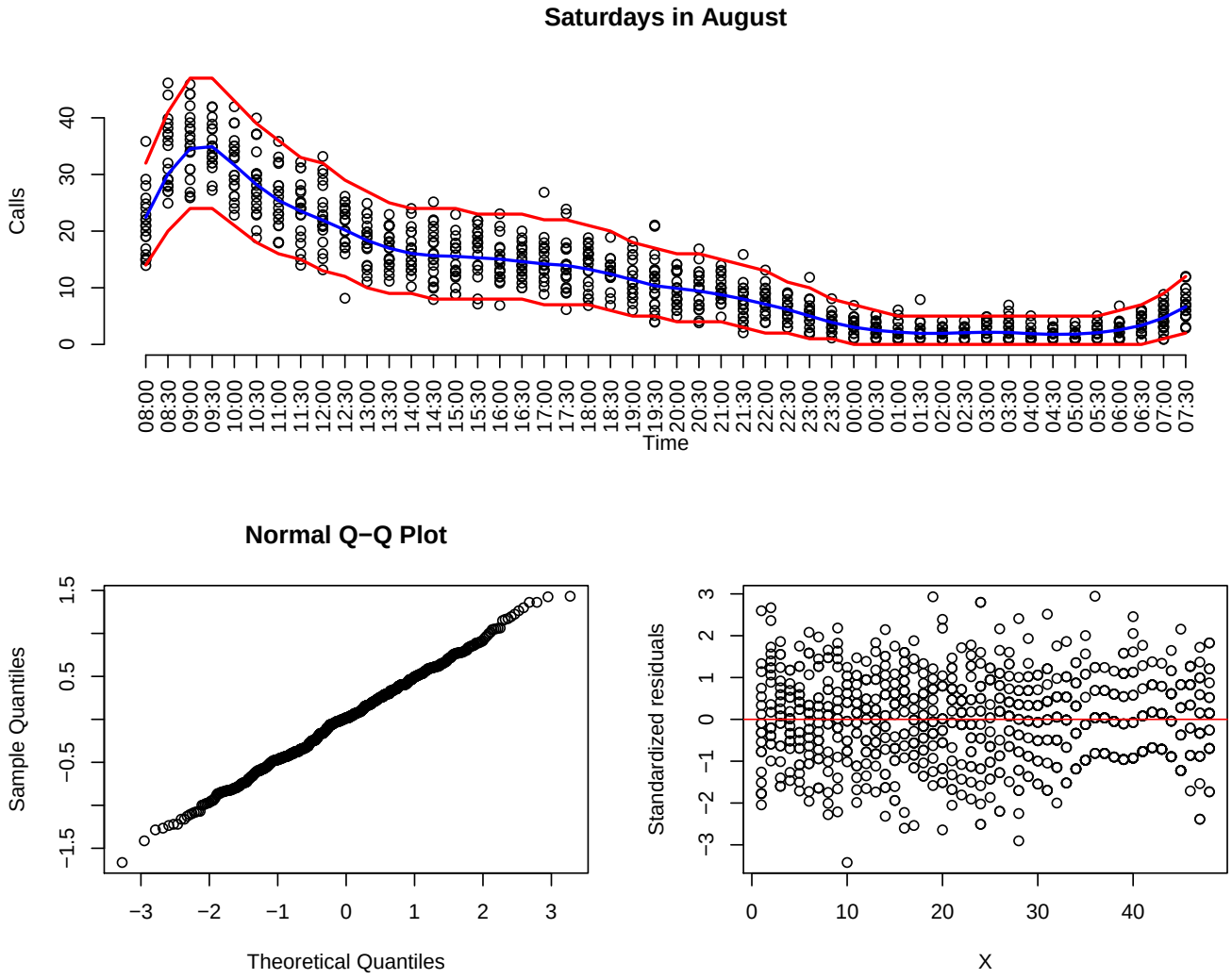

Figure 8: Observations with loess fit, Poisson-based errors, and 95% prediction interval.

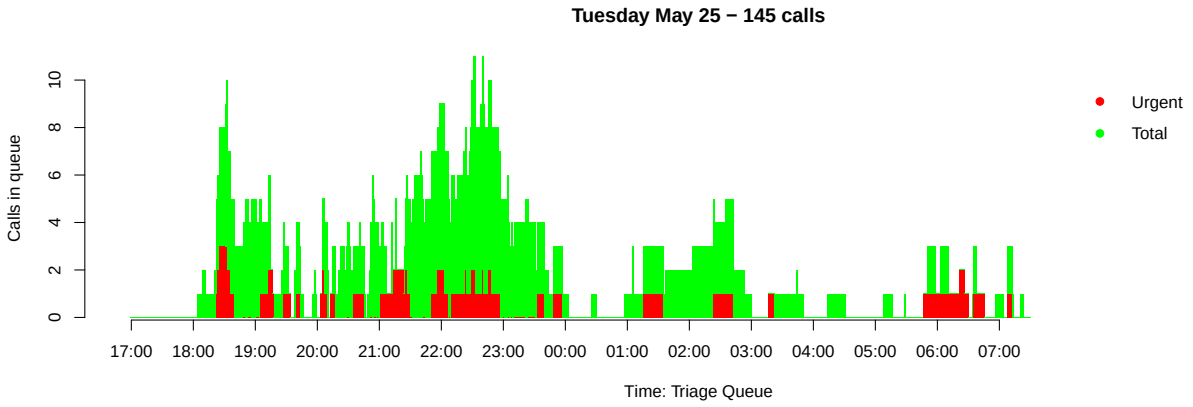

Figure 9: Triage queue plot

5 Operational delivery given forecasts

The main part of this paper has described how we may obtain good interday and intraday forecasts of call arrivals. Given these forecasts we now discuss some of the possibilities in using them.

5.1 Monitoring queues and staff service times

Call centre managers need quick effective tools to monitor performance and diagnose problems. With regard to the service rates of staff handling patients, these will vary according to patient needs and because of differences between staff. In order to maximize the value of the service offered by the organization, there is a need to identify good and poor practice in staff. This is difficult to assess for medical call centres because we have no means of identifying whether a treatment offered to a patient was appropriate to their illness. Ideally we would trace patients through the full system and try to relate patient outcome (appropriateness of treatment) to the pathway followed; however this is not yet feasible. Instead we have only the amount of time each GP took to triage each patient. It is not necessarily true that short triage times are preferred to long service times, as a long triage time may properly reflect the level of treatment required. However, it may be reasonable to assume that, given sufficient numbers, the caseloads assigned to different staff members may be representative of the underlying caseload. Under this assumption, we compare the service times of GPs.

All out-of-hours cases were collected from January 2007 until 24^{th} November 2009. From these, 250,000 telephone consultations were recorded and classified by clinician. Service times were modelled adjusting for urgency (urgent cases typically take longer) and case outcome (for example, home visits need to be arranged). There is some clear dependency of service time on calendar or shift effects. Typically GPs tend to triage more quickly when there are patients in the queue, and this tends to be the case when staffing does not appropriately reflect intraday and interday forecasts. An example of the build up of the triage queue is shown in Figure 9, a typical diagnostic we may offer managers. However, we do not wish to build low/high caseload features into a model. We will assume that service times broadly follow an exponential distribution, in which case a generalized linear model employing Gamma family errors (log link) is suitable. Our focus is then on residuals from this model, summarized by clinician.

Figure 10 shows a boxplot of differences from the mean service time, classified by clinician, for weekday evenings in the second quarter of 2010. Each service time has been adjusted for case type and outcome. Eight 'quicker' GPs appear in the left of the graph and five 'slower' GPs appear in the right. All these GPs had 75% of triages below (above) the adjusted average. 33 GPs behaving similarly to the adjusted average are omitted from the picture as we wish only to identify aberrant behaviour. We restrict model fitting to those clinicians dealing with at least 20 patients. The box widths are proportional to size of caseload per GP. Observe that some of the faster GPs tend to have larger caseloads, suggesting that experience is relevant to service time. Such analyses are invaluable to managers. They quickly identify differences

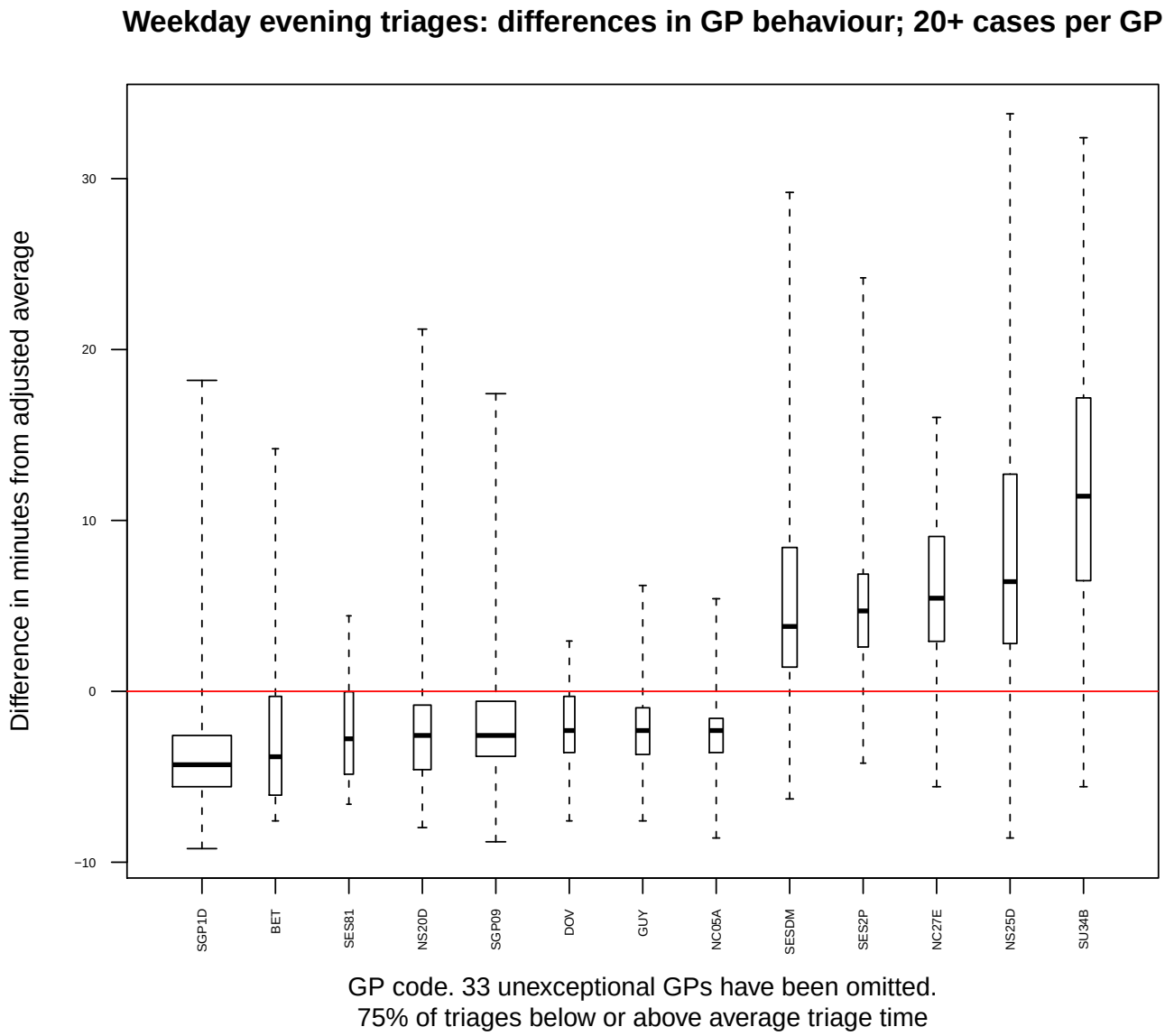

Figure 10: Comparing GP triage lengths: Weekday evenings

between members of staff. The interpretation made in this example is that faster GPs tend to be those with most experience. This lead to enhanced training for new members of staff and review procedures for some other staff. In addition, the organization adjusted their staff shifts in order to match experienced and inexperienced staff. The features revealed in this analysis correlated quite closely with the organization's prior judgements.

5.2 Shift Allocation

Given good forecasts of call volumes, it is relatively simple to use simulation techniques to establish the number of GPs that need to be on duty. The triaging queue that we describe is an example of a non-preemptive priority queue with two priority classes (routine and urgent) and different service rates. See, for example, D. Gross and Harris (2008) for analytic rate-balance equations for this scenario. The ingredients to the simulation study are service time distributions for routine and urgent triages, and the arrival rates for each type of patient. From data, service times for patients classified as urgent do not seem to vary during the day, which is what we would expect. Service times for patients classified as routine appear to be slightly shorter at busier times of the day, suggesting perhaps that GPs take a more relaxed approach

during off-peak periods. We don't initially want to build such a feature into the simulation and so take service rates as averages across the data set. Note that there are subtle issues involved here. Clearly one could set the service rate at that which GPs could achieve during peak periods, but is it desirable to require staff to be operating at peak stress throughout a shift?

For this data set, lengths of service times are taken to be exponentially distributed with mean 543 (582) seconds for routine (urgent) triages. The other ingredients to the simulation are the arrival rates. These vary by time of day and calendar day, as modelled in earlier sections. Therefore, we must run a separate simulation for each different kind of calendar day. In section 4 we showed how to estimate arrival rate of calls during each half-hour period. We now suppose more formally that arrival rates during a day follow a nonhomogenous Poisson process with arrival rates given by our estimates. There are 48 half-hourly periods with estimated arrival rates; two examples are shown in Figure 11. In order to simulate from this process, we use the procedure described in Ross (1985); see Goegebeur (2011) for a summary algorithm. Thus, imagine arrival rates $\lambda(1), \ldots, \lambda(48)$. To simulate from this process, we choose λ such that $\lambda(t) \leq \lambda$ for all $t \leq T = 48$. We then generate events according to a Poisson process with rate λ, and accept an event at time t, with probability $\lambda(t)/\lambda$ and independently of preceding decisions. The process of counted events then constitutes a nonhomogenous Poisson process with the required intensities.

Given simulated patterns of call arrivals during a day, we may explore the implications of different allocations of resource (GPs) to handle the calls. The driving criterion is satisfaction of KPIs, for example that the number of breaches of triage waiting times for patients is no greater than 5% of call volume. For our example, the cost of idle resource is irrelevant insofar as patient treatment within reasonable time is all that matters. However, for other scenarios, for example call centres for banks, we may wish to optimize according to both KPI targets and to costs of idle call handlers. The optimization itself is constrained by allowable shift patterns. Typically GPs work a fixed number of hours, essentially continuously. Further, shift solutions are expected to align with customary starting and ending times. The constraints set by the collaborating institution are that the minimum (maximum) shift lengths are 4 (8) hours. There are a number of other constraints to be taken into account. This problem falls into a general class of hard problems known as the nurse scheduling problem; for a recent detailed approach see Parr and Thompson (2007).

However, the following heuristic offers a reasonable solution. To find a starting solution, note that theoretical analysis Ross (1985) requires that the utilization factor of the servers, $\rho = \lambda/(c\mu) < 1$, where c is the number of servers, λ the mean arrival rate, and μ the mean service rate. As for each half hour we have reasonable estimates of μ, λ, this suggests a starting value for the number, c, of GPs required in each half-hour interval to achieve better than equilibrium. Once a starting solution has been found, and once that solution has been modified to suit shift constraints, it is reasonably simple to add or withdraw shifts and assess the effect on targets.

Further related simulation studies are required for, for example, allocation of emergency vehicle drivers for GPs assigned to home visits. We noted earlier that GPs on their way to or from a home visit may help remotely with the triage queue; data on such piecemeal triaging needs to be collected and the effects on queue lengths can then be simulated as periodic extra resource. One of the outputs from the simulation study is a better understanding of some of the uncertainties inherent. Given sufficient simulations, we can produce an estimated probability distribution for various outcomes: number of breaches of targets, and so forth. In particular it is very easy to model the effects of local shocks to the system. It is well known that small unexpected surges in demand can have severe effects on time of wait in queue as the number of servers struggle to catch up. For this scenario, we can use simulation to explore queue build and triggers to bring extra servers online to cope. Of course, such simulation strategies have a long history in this area, for example see Banks et al. (2010).

6 Conclusion

In this paper we have described statistical approaches to a number of complicated and interrelated issues involved in planning, forecasting, and monitoring for an out-of-hours medical call centre. We have provided methods to forecast daily call volumes and the pattern of calls arriving during the day. These work very well and are now embedded within the practice of our collaborating organization. These straightforward

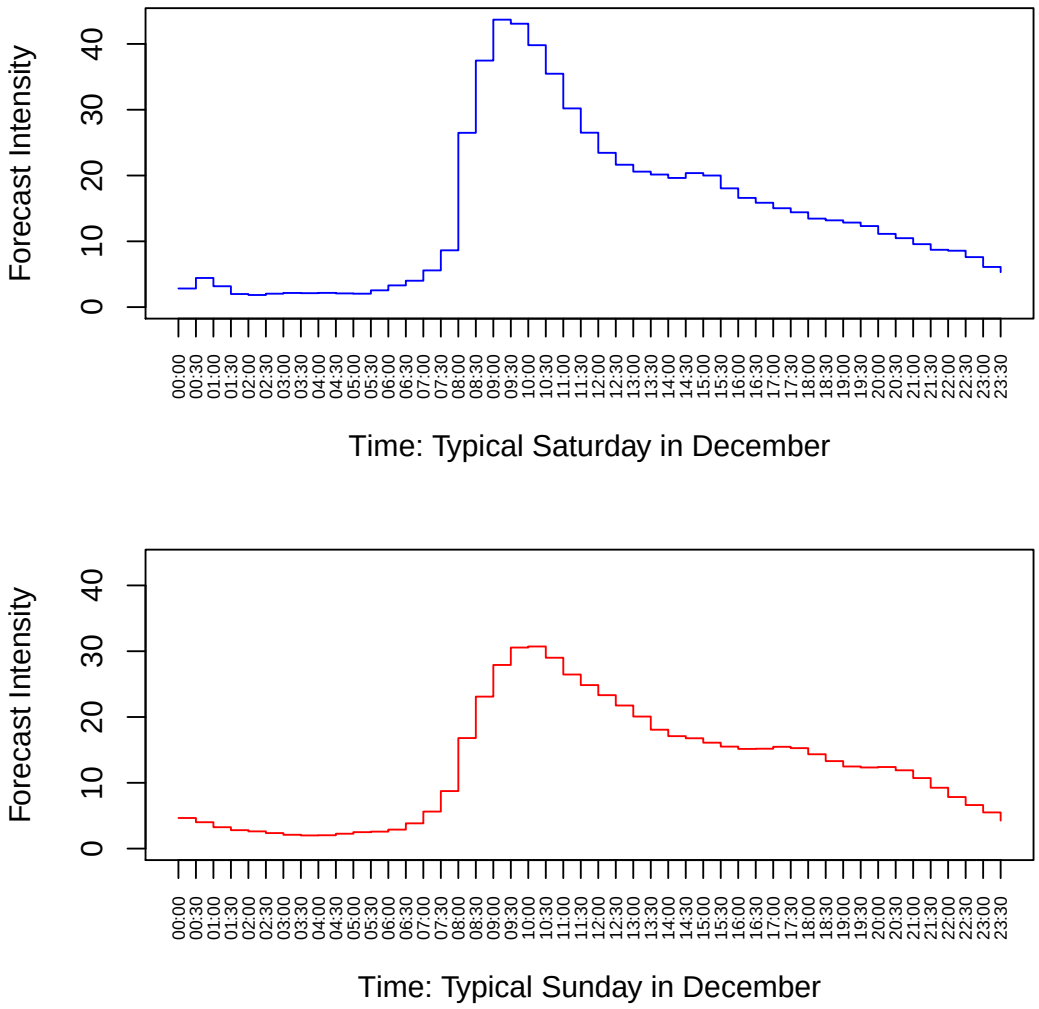

Figure 11: Estimates of arrival rates per half-hour for typical Saturdays and Sundays each December.

methods capture a high proportion of explainable variation. We have provided methods showing how we may begin to monitor staff activity having adjusted for relevant factors. We show how simulation methods may build on inputs from the forecasts to allocate staff resource. It is relatively trivial to adapt the methods to other kinds of call-centre.

We have not yet linked interday and intraday forecasting statistically. As we noted, we normalize our intraday forecasts to the volumes suggested by the interday modelling, and this ignores discrepancies that may arise between them, and fails to take account of correlations between them. From the operational perspective, this is a reasonable feasible solution. KPIs tend to focus on daily activity such as call volumes and queue lengths (interday), whereas for planning work rotas and shifts it is more useful to forecast arrival rates over a shift (intraday).

It is worth describing the intellectual resource required in implementing this methodology. One requirement is the availability of historical data allowing sufficient degrees of freedom to estimate regression model parameters. Longer historical time periods permit finer details to be captured by the modelling. Cleaning the data often exposes problems in its original collection, requires resolution of ambiguities, and so forth. These are themselves valuable from the institution's perspective and tend to lead to improved data collection processes. The basic modelling described in section 3 and section 4 has to be carried out by a competent statistician, who will also be responsible for addressing resource allocation via simulation. Once these have been carried out, software can provide routine forecasts and diagnostics and these can be managed by junior staff without day-to-day supervision from a statistician. For our work, we used the **R** package R Development Core Team (2009) for the background modelling and for provision of daily forecasts, and a Microsoft Excel front end Neuwirth et al. (2010) to present all outputs in a format acceptable to our collaborating institution, and which could be readily tied in with their KPI systems. It is necessary to maintain contact with the statistician for several reasons. The two most important are to monitor performance and diagnose discrepancies in model behaviour, and to handle extra modelling when there are external shocks to the system, such as unforeseen epidemics. This is critical because otherwise the value of the models decays and institution staff begin to lose trust in the forecasts.

7 acknowledgements

Some of this work was jointly funded by One North East and the Economic & Social Research Council under Knowledge Transfer Partnership 6732. We are grateful to Northern Doctors Urgent Care Group and their staff for providing the data and background insights. Some of this work was presented to the 73rd annual conference of the Institute of Mathematical Statistics, 9th-13th August 2010, Gothenburg. Some other aspects were also presented as part of a keynote talk to the YOR17 conference, 5th-7th April 2011, Nottingham. Some example simulations of GP shift patterns were carried out by K.R. Ford, a Durham University M.Math student.

References

Avramidis, A. N., A. Deslauriers, and P. L'Ecuyer (2004). Modelling daily arrivals to a telephone call center. *Management Science 50*(7), 896–908.

Banks, J., J. S. Carson, B. L. Nelson, and D. M. Nicol (2010). *Discrete-event system simulation.* Upper Saddle River, NJ: Pearson.

Chambers, J. M. and T. J. Hastie (1993). *Statistical models in S.* London: Chapman & Hall.

Cleveland, W. S. (1979). Robust locally weighted regression and smoothing scatterplots. *Journal of the American Statistical Association 74*(368), 829–836.

Cleveland, W. S. and S. J. Devlin (1988). Locally weighted regression: An approach to regression analysis by local fitting. *Journal of the American Statistical Association 83*(403), 596–610.

D. Gross, J.F. Shortle, J. T. and C. Harris (2008). *Fundamentals of Queueing Theory.* New York: Wiley.

Draper, N. R. and H. Smith (1998). *Applied Regression Analysis*. New York: John Wiley & Sons.

Goegebeur, Y. (2011). The poisson and the nonhomogeneous poisson process.

Hand, D. J. (2006). Classifier technology and the illusion of progress (with discussion). *Statistical Science 21*(1), 1–34.

Ladiray, D. and B. Quenneville (2001). *Seasonal Adjustment with the X-11 Method (Lecture Notes in Statistics; 158)*. New York: Springer.

Neuwirth, E., with contributions by R. Heiberger, C. Ritter, J. K. Pieterse, and J. Volkering (2010). *RExcelInstaller: Integration of R and Excel, (use R in Excel, read/write XLS files)*. R package version 3.0-19.

Parr, D. and J. M. Thompson (2007). Solving the multi-objective nurse scheduling problem with a weighted cost function. *Annals of Operations Research 155*(1), 279–288.

R Development Core Team (2009). *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing. ISBN 3-900051-07-0.

Ramsey, J. O. and B. W. Silverman (2005). *Functional Data Analysis*. New York: Springer.

Ross, S. M. (1985). *Introduction to Probability Models*. Orlando, Florida: Academic Press.

Shen, H. and J. Z. Huang (2005). Analysis of call center arrival data using singular value decomposition. *Applied Stochastic Models in Business and Industry 21*, 251–263.

Shen, H. and J. Z. Huang (2008). Interday forecasting and intraday updating of call center arrivals. *Manufacturing and Service Operations Management 10*(3), 391–410.

Stirling, S. G. (2011). Statistical methods for supporting urgent care delivery. Thesis, Durham University.

Weinberg, J., L. D. Brown, and J. R. Stroud (2007). Bayesian forecasting of an inhomogeneous poisson process with applications to call center data. *Journal of the American Statistical Association 102*(480), 1185–1198.